# 1 Introduction

History refers to the late 19th century as the industrial revolution. Our era is certain to be known as the information revolution. Information technology profoundly impacts our everyday life, from mobile and satellite communications to computers and the Internet. We are growing increasingly dependent on fast, reliable, and secure processing of information.

The information revolution has raised the stakes in terms of operational standards and demands, establishing a new range of technical challenges. The rapid advances in reliable transmission, storage, and processing of information are often driven by the development of improved algorithms for computationally hard problems [1]. For example, the recent increase in hard drive capacity was facilitated by efficient error-correction algorithms that are built directly into the hardware.

However, much more will be needed to maintain this pace. Practical algorithms have typically been developed on a heuristic level, without a theoretical basis, which severely limits their scope of applicability. To fill this gap between engineering needs and theoretical computer science, we propose an innovative physics-inspired research and development program for the *analysis and development of new algorithms*. Our goal is to reduce the computing time and memory space required for complex operations in information processing.

The time is ripe for this research as recent developments illustrate the utility of the proposed research. Error probabilities in efficient, nearly optimal, error-correcting codes [2, 3] can be quantified through a remarkable mapping [4, 5] to statistical physics models of disordered magnets [6, 7, 8]. There is significant progress in understanding the computational complexity of generic tasks, a central problem in computer science [9, 10]. Significantly, the threshold separating computationally tractable and computationally intractable tasks [10], as well as the threshold separating correctable and incorrectable domains in error correction [5], are both closely related to the phase transition between order and disorder in magnets. The powerful theoretical framework of critical phenomena, including the concepts of scaling and universality [11], thus leads to the creation of advanced techniques that vastly outperform existing methods in the information sciences.

## Institutional Goals

There is a need for the United States to maintain its leadership in the field of information technology with economic and national security implications. We cannot depend on foreign countries to provide the strategic technologies and knowledge base critically needed in defense and national security computing applications. Our proposal is one step to develop a long-term capability at Los Alamos in this emerging area of information science.

Indeed, the development of efficient algorithmic solutions for computationally hard tasks is fundamental to the Laboratory's mission and impacts programs in modeling and simulation, verification and validation, nonproliferation, and infrastructure security. The underlying real-world problems are complex and enormous efforts are required to decompose them into a subset of tasks that are computationally tractable and feasible for analysis. These problems are ubiquitous in science-based prediction for complex systems, and impact our predictive and simulation capabilities. They are at the core of the strategic areas identified in our LDRD/DR white paper as high priorities for computer and computational science.

Our proposed research fits into a long-standing tradition at the Laboratory. From MANIAC to the introduction of the Monte Carlo method, to the invention of the scholarly information archive (xxx.lanl.gov), there has been a lively institutional dynamic that involves leveraging our strengths in the physical sciences to secure leadership in information and computer science. Los Alamos has made a strategic decision to invest in theoretical statistical physics for national security applications, and this investment is paying off with visionary programmatic work, high-impact basic research, and recruitment of outstanding young talent.

Our team includes theoretical physicists and information and computer scientists, who are already working in close collaboration. This interdisciplinary group is highly visible in the international research community. Our team has a strong track record in analyzing and developing algorithms using physical principles. We were the first to apply theoretical physics methods including field theory for the statistical analysis of data transmission fidelity in fiber optics communications [13] and for quantifying error probabilities in error-correcting algorithms [15]. We pioneered the use of traveling waves for the analysis of data storage, compression and zipping algorithms [16, 17]. Our team has also developed a superior novel community detection algorithm that combines a mapping to interacting particles with molecular dynamics methods [18] and the Extremal Optimization algorithm used successfully on notoriously intractable optimization problems [19].

# 2 Research Goals and Objectives

This research concerns a set of inter-related problems that share a common theme: they can be described in statistical physics terms as large discrete systems of interacting "spins". Our objective is thus the development and analysis of efficient algorithms for these problems.

The performance of an algorithm is quantified by the amount of memory space and computer time it requires to perform the prescribed task, as well as by the quality of the solution found. Algorithms that always find the best solution to computationally hard problems require exponentially large time, while algorithms that find near-optimal solutions can run much faster. We will consider both the average performance of an algorithm and the worst-case performance in extreme scenarios. The quality of the solution in the worst-case is especially important in situations where high reliability is required.

We will develop methods that work well in practice. Data sets on personal computers can easily involve $10^{12}$ bits, while supercomputing and bank applications involve even larger data sets. The techniques of theoretical physics, developed to deal with macroscopic systems of $10^{23}$ particles, are thus a natural way to deal with such sizes. Moreover, theoretical physics has a well-developed methodology for describing problems of large but finite size. These issues are becoming especially important in modern physics applications at the nanoscale and are now being imported into information science where finite-size performance is obviously crucial.

The impact will be at both the fundamental and applied level. In computer science, we will address complexity and NP-completeness, the most basic problems in the field. The problem of NP-completeness is one of the seven Millennium Problems posed by the Clay Mathematical Institute as the most important problems in mathematics today. Work in theoretical computer science has proven of great practical importance: *there are examples where today's algorithms, running on a computer from the 1950s, would outperform the 1950s algorithm running on a modern computer*. Applications of modern algorithms have impacted hardware developments in the fields of data storage and transmission that we consider. The application of physics techniques to computer science is opening a new subfield which LANL must actively pursue to stay in the forefront of information technology.

## 2.1 Algorithm analysis

We will exploit powerful recent developments at the interface of theoretical physics, information theory, and computer science to analyze the performance of existing algorithms, such as the *belief propagation*[1] algorithm that is broadly used in the three fields. We will study algorithms for data transmission problems in information science where error correction is a central challenge. In addition, we will analyze fundamental combinatorial issues of computer science that arise in verification and validation problems.

The best approximation algorithms find solutions that are close to optimal in all but a few cases. For example, the *belief propagation* algorithm fails when loops in the underlying graphical structures become important. We will analyze these cases using rare event and extreme deviation theory. Additionally, we will use phase transition methods such as scaling, renormalization, and universality to characterize algorithms that work well in certain regimes but perform poorly in others where the problem becomes intractable.

## 2.2 Algorithm development

Many information science problems can be mapped onto physical interacting particles systems including magnetic and glassy systems. Our research and development approach is to find such mappings and then utilize them for the design of new algorithms.

We will develop heuristic algorithms – methods that work well in practice but do not necessarily guarantee optimality. To improve performance, we will recast existing heuristic algorithms as the first step in a series of systematically improvable approximations. The theoretical analysis discussed above will be used to quantify this improvement.

The proposed research detailed below is organized into two categories: **information science** and **computer science**. We stress that this division is somewhat artificial since there are many practical connections and intellectual ties between the various projects.

# 3 Proposed research: Algorithms in Information Science

Our focus is processing, coding, and retrieval of information with emphasis on characterization of data corruption. Data corruption may occur in transmission of data through communication lines, and in retrieval from storage devices. In the former case, physical imperfections such as disorder and noise in fiber optics channels cause corruption, while in the latter case interference from neighboring symbols and instrumental noise cause the head to misread stored information. We will perform a comprehensive statistical analysis of performance characteristics of transmission lines and storage media. This information will be used to generate improved data transmission and restoration algorithms.

Another major theme is error-correction. To overcome data corruption, redundant transmission of information is used. We will analyze performance of recently developed error-correction algorithms. We will quantify performance of error-correction schemes by analyzing their dependence

[1]**Historical note:** *Belief propagation* is ultimately related to an approximation introduced by Hans Bethe in 1935 to explain melting of lattices [12]. In an attempt at mathematical simplification, Bethe considered a tree structure instead of a regular lattice. The solution of a statistical model on a tree has since been called the Bethe tree approach. Robert Gallagher, who invented the belief propagation algorithm in coding theory [2], was unaware of the physical Bethe tree approach but, nevertheless, constructed an algorithm based on exactly the same idea: ignoring effects of loops in the graphical model representing an error-correction code. Gallagher's constructive idea was to consider the family of codes, called Low-Density-Parity-Check (LDPC) codes, with as few short loops as possible.

on the signal-to-noise ratio characteristics of coding and decoding. We have recently developed a comprehensive theoretical framework (instanton analysis) for describing the performance of error correcting codes [15] and we will use this framework to design improved coding schemes. The improvements promise a qualitative, orders of magnitude, increase in efficiency.

## 3.1 Data transmission

Development of high speed, broad bandwidth, photonics for optical fiber communications is facing major limitations both in signal processing and signal transmission. Both material irregularities and environmental noise can destroy optical pulse integrity and cause information loss. Modern standards are high: typical Bit-Error-Rates (BER), or probability of error per bit, should not exceed $10^{-12}$ and thus, there is a need for analytical methods, particularly probabilistic analysis for quantification of information loss in transmission processes.

We will develop theoretical and computational tools for quantifying the performance of fiber-optics channels. For a given realization of the disorder due to fiber birefringence, fluctuations in chromatic dispersion, or microscopic material disorder, we will estimate the error probability due to instrumental noise generated by the laser source or the amplifier. We will deliver quantitative descriptions of performance fidelity of realistic photonics-based communication systems extending our previous research [13]. This analysis will be used to calibrate engineering tests and actual experiments in optics communications [14]. This will be invaluable in designing communication systems, enabling us to predict performance before doing the fabrication. We will answer the basic quality assurance question: does the BER meet the tight quality requirements given the physical parameters of the fiber-optics system?

## 3.2 Data storage and retrieval

Data storage and retrieval is a classic area of information science [20]. Handling massive amounts of data requires compression and then decompression of data and storage and then retrieval of data. We propose to develop new theoretical methods for first characterizing and than qualitatively improving the performance of data storage and retrieval algorithms.

Many data storage and sorting algorithms utilize tree architectures. The novelty of our approach is mapping trees to collision processes in a gas and utilizing nonlinear dynamics and stochastic processes techniques. We have already used this approach to characterize the best and the worst case scenarios for standard data storage and retrieval algorithms [16, 17]. Our goal is to address how the retrieval procedure scales with the size of the data set. Knowing the scaling should allow us to suggest more efficient storage and retrieval schemes. We propose to use nonlinear waves techniques to analyze the best and the worst case scenarios. Our preliminary work shows that an important characteristic, the rank of the tree, can be described using this approach. In parallel, we will develop new Monte Carlo algorithms for efficient generation of large scale tree structures.

In addition, we will also address data retrieval processes. Because recording densities of conventional data storage technologies are approaching a plateau, novel technologies are emerging. Much of the current effort is dedicated to new technologies that utilize natural architectures such as a *two-dimensional lattice*, instead of the traditional co-centric (e.g., reading bits off a spinning disk) architecture [21]. Similarly, holographic recording may lead to three dimensional architectures. This emerging technology presents a challenge: overcoming data corruption that may occur in the reading process because the output for an individual bit may be polluted by signals from neighboring bits. The task at hand, data restoration, presents a nontrivial inference problem; *maximal likelihood*, the exhaustive analysis of all configurations to find the most likely one, is not practical. Using theoretical methods, developed in the physics of disordered magnets, we will map the multi-dimensional inference problem to a spin system. Perfect data restoration requires finding the lowest energy state of the magnet. This mapping will be used to develop efficient algorithmic solutions for data restoration. We will utilize the mean-field approximation when the interference is long-ranged. Otherwise, we will develop a systematic approximation where nearest-neighbor bits are considered at the first level, next nearest-neighbors at the second level, etc.

## 3.3 Error correction

The most basic error-correction technique is to repeat each unit of data multiple times, just as one might need to repeat a sentence several times in order to be understood over a poor telephone connection. However, it is possible to detect and correct errors with a far less redundant approach. The best error-correction method in a given case depends on the statistical characteristics of the channel noise. A general coding scheme may be understood as follows. The transmitter wishes to send data over a noisy channel and so encodes the data in a redundant form, sending a longer message over the channel. The receiver gets a corrupted form of the message, and then tries to reconstruct the transmitter's original message. *Maximal likelihood* decoding generically becomes intractable for codes with even tens of bits. Therefore, the practical decoding challenge is to reconstruct the message with minimal computational complexity.

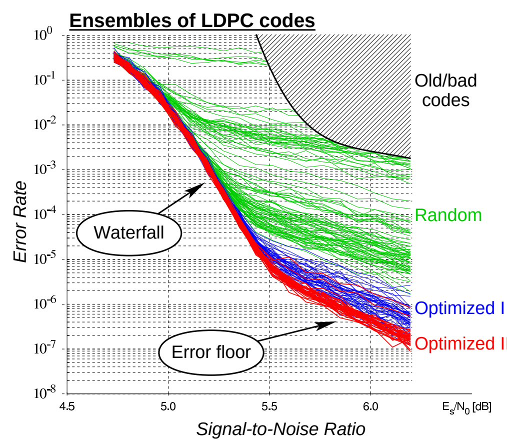


Figure 1: Performance of ensembles of LDPC codes.

A novel exciting era has started in coding theory with the discovery of LDPC [2] and the so-called turbo [3] codes. These codes are special, not just because they can approach very close to the virtually error-free Shannon transmission limit, but mainly because a computationally efficient *belief propagation* algorithm exists. (While the computational time of the *maximum likelihood* algorithm scales exponentially with the number of bits, $\sim 2^N$, the scaling is linear for the *belief propagation* algorithm, $\sim N$.) These superior codes are widely expected to replace older codes in communication and storage within a few years [1]. In Fig. 1, we show the performance for various LDPC codes, using the belief propagation heuristic for decoding. The error rate decreases with increasing signal-to-noise ratio along the channel. The decrease in error rate is rapid in the so-called waterfall region, but a much slower decrease in error rate is seen in the error floor region. The error floor is a result of the heuristic decoding used. Although maximal likelihood performs much better in this region it is prohibitively expensive for large numbers of bits. The different curves show a sequence of improvements to the code which indicates the possibility for further improvements.

The analysis of these codes is difficult. In optical communication systems, the maximum acceptable probability of an error per bit, BER, is $10^{-12}$. For hard drive systems in personal computers $10^{-15}$ is the maximum acceptable BER and $10^{-20}$ is the maximum acceptable for storage systems used in banks and financial institutions. However, brute force numerical methods, such as Monte Carlo, are not feasible for BER below $10^{-9}$.

We will develop methods for analyzing the domain of very weak noise. In our physics approach [15], we recast the problem of estimating low error probabilities as a computationally and theoretically tractable minimization problem aiming to find the most probable noise configuration, the instanton. An instanton configuration of noise found numerically is shown in the left panel of Fig. 2 for a particularly popular code decoded by the *belief propagation* algorithm. On the right panel of Fig. 2 this instanton prediction for the error rate is compared to Monte Carlo simulation results, showing a good quantitative agreement in the error-floor domain for the proof of principles test. Note that *there exist no other methods capable of such error-floor analysis*. We will follow these promising results and use the instanton method to investigate the dependence of the error-floor asymptotics on the coding scheme, number of bits, and type of the communication channel. We will also establish a connection of this instanton approach with the linear programming approach of [22]. This study will lead to creation of a universal computational toolbox providing efficient performance diagnosis for coding and decoding schemes. We will use this toolbox for testing new algorithms. For example, one algorithmic idea that we plan to explore is to find a sequence of approximations to the ideal, but impractical, maximum likelihood algorithm, such that belief propagation algorithm is the first one in the sequence. The higher the order of an algorithm in the sequence the better its decoding quality is expected to be while the respective computational complexity will still be kept polynomial. We expect that the unique capability in code analysis and design resulting from this research will be very valuable as the use of LDPC codes becomes more widespread [1].

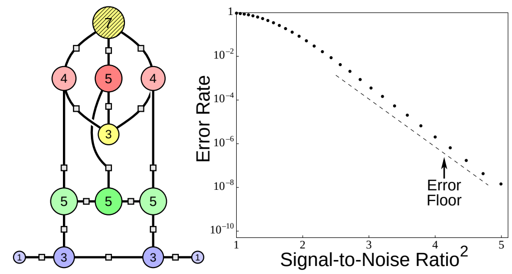


Figure 2: Left: instanton configuration of noise. Right: theoretical prediction against simulation results.

# 4 Proposed research: Algorithms in Computer Science

We will focus on information processing, with an emphasis on combinatorial optimization. Combinatorial optimization concerns finding the best solution from a set of discrete alternatives. The above problem of finding the most likely initial message in decoding can be rephrased in a language of combinatorial optimization. We will study two important computer science problems, satisfiability and community detection.

## 4.1 Satisfiability

Satisfiability is one of the canonical problems of computer science. Given a set of logical clauses and values must be assigned to Boolean variables to satisfy the maximal number of clauses. An exhaustive analysis of all cases is not practical, but certain heuristic algorithms can work very well for typical situations. Physics methods have proven very powerful both in analyzing and designing these algorithms. Cases in which the heuristic algorithms run slowly have been shown to be associated with a phase transition in which the collective behavior of the system causes the optimization problem to become intractable. ( See Figure 3.) The analysis of the phase structure in terms of spin-glass theory has led to the development of the powerful *survey propagation* algorithm. This algorithm generalizes the *belief propagation algorithm* and thus, is also useful in the context of error-correction.

The problem of finding a satisfying assignment to a set of clauses is demanding. Technically, it is NP-complete: if one could solve this problem in polynomial rather than exponentially long time, one could also solve a broad class of other problems in polynomial time. There is no known way of doing this, and it is strongly believed that none exists: the problem is "intractable". An exciting recent development has been the understanding that the intractability of the problem is limited to certain special cases: there are many regimes in which it can be solved rapidly. Physics methods have been used to associate these cases with phase transitions, to predict their properties and, most significantly, to create highly effective algorithms for solving the problem.

Solving instances of satisfiability is also a very practical problem: it is used in *formal verification* [23] to rigorously prove the correctness of computer programs (Microsoft has an outstanding research group applying statistical physics and computer science to formal verification). Based on our preliminary results [24], we will investigate the performance of satisfiability algorithms in the context of NuSMV, a widely used program for formal verification, and we will develop algorithms for formal verification of multiagent simulations of sociotechnical systems used in LANL projects, such as AdHopNet.

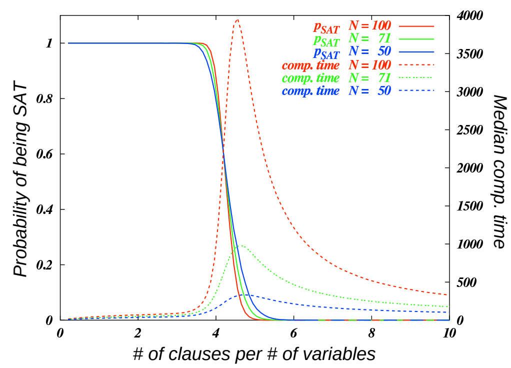


Figure 3: Phase transition in satisfiability. The solid curves show the solutions becoming rare near the phase transition, while the dashed curves show a peak in the computational resources required to find the solution.

Our next task is to provide an analysis of *survey propagation*. This is a remarkable algorithm that, based on knowledge of the problem's phase structure brought about through spin glass theory, has increased by two orders of magnitude the size of satisfiability instances that one can solve in practice [10].

We will then analyze *Approximation Algorithms* [25], algorithms yielding solutions that, while not optimal, are quantifiably close to optimal. The analysis of such algorithms in the past has largely focused on their worst-case behavior, and has not substantially considered the typical-case complexity of obtaining good solutions. However, the analysis of *survey propagation* may be extended to approximation algorithms, clarifying precisely this typical-case scenario. We will thus study the extension of methods inspired by statistical physics to approximation algorithms.

We will also use the methods of scaling, universality, and renormalization group to study the situation near phase transitions in these systems. Phase transitions are generic to many combinatorial problems and are essential to analyzing the most difficult cases of these problems. To date, there remain a number of interesting combinatorial problems whose phase transition is not well understood, such as *graph bipartitioning* [26], a problem with many practical applications in resource allocation as well as hardware design. We have started to investigate the phase transition in graph bipartitioning by identifying the relevant order parameter [27], but the order of the phase transition has not been determined conclusively. We intend to determine it, enabling a host of results from critical phenomena to be applied to algorithmic improvement.

A complementary approach is the use of kinetic theory to understand these phase transitions. The phase transition is due to the emergence of collective structures, where the change in a single variable's assignment can require many other variables to be reassigned to satisfy the clauses. In its simplest form, the problem reduces to the emergence of cycles in random graphs [28]. We have recently developed a comprehensive kinetic theory of structural properties of such structures [29], capable of characterizing the nature of the satisfiable-to-unsatisfiable transition. We propose an ambitious investigation, characterizing the transition in more complicated cases that require an analysis of the topology of complex manifolds, such as tori in random membranes.

Finally, we will provide an analysis of the rare difficult instances that are far from the phase transition. In this regime, methods such as belief propagation or survey propagation work well for almost all instances, but still occasionally fail in a few instances. We will perform an instanton analysis, reminiscent of [15], for evaluating the structural origins of the rare unsatisfiable events occurring in a typically satisfiable phase. Knowledge of these instances is important for testing and design of new algorithms.

## 4.2 Clique and Community detection

These are both problems of finding clusters of nodes in a given graph that are closely coupled to each other. Clique detection is another NP-complete problem, that of finding the largest clique, or group of nodes, that are all connected to each other in a graph. We attack this with a novel in-house developed algorithm [18], called the Simplex Distortion Algorithm (SDA). Community detection is the problem of finding the optimal assignment of communities in a graph, which we attack by casting it as a probabilistic inference problem.

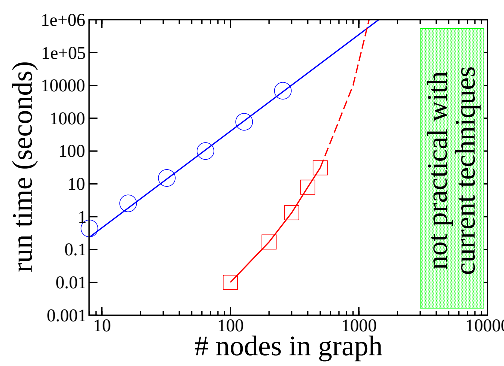

Figure 4: Runtime of SDA (circles) and standard algorithms (squares).

The central idea of our approach to finding cliques, the Simplex Distortion approach, is to recast combinatorial problems as a high dimensional physics problem. We map each node of the graph onto a physical point. The evolution will draw together those nodes that are connected, and push apart those that are disconnected. This maps the abstract "clusters in a graph" into geometric clusters in Euclidean space which can be readily identified. We plan to experiment with clusters of various sizes, degree of internal connectedness, and "noise" levels from exterior connection into which the clusters are embedded. In the pilot research, we have found empirically that the original clusters "hidden" via random vertex re-labeling can readily be retrieved. By tuning the strength of the repulsion between unconnected vertices, we can select more strongly connected clusters, eventually culminating in perfect cliques. The test shows (see Figure 4) that for random clusters, this approach appears to have approximately $N^3$ polynomial scaling in runtime over a range of problem sizes. We expect that the SDA algorithm will surpass the "state of the art" algorithm [30] for very large graphs.

Community detection is the problem of breaking a graph into communities, such that nodes in a community are highly connected to each other, but not to the outside [31]. The problem is relevant to graph partitioning problems, as well as to graphs found in social or biological systems. Most community detection algorithms define communities in an ad hoc way. We instead suggest the following procedure inspired by procedures commonly used to evaluate community detection algorithms: we consider a given initial assignment of nodes to communities, and we consider a graph in which only nodes within the same community are connected. Then we imagine that some of the connections are re-wired with a given probability so that nodes within different communities will be coupled. We then cast the problem as an inference problem by picking the community assignment which has the maximal likelihood of leading to the final graph. We expect that if the number of communities is small compared to the number of nodes, belief and survey propagation algorithms will be effective. The systematic basis of this approach is expected to provide better performance than existing ad hoc methods, while the high speed offered by belief propagation algorithms will be essential in dealing with large data sets in biological or social systems.

# 5 Schedule of Deliverables and Milestones

**FY06:**

• Derive realistic models for optics channels. Develop a mean-field method of decoding for the 2d and the 3d ISI interference problem [MC,MS]. Develop new nonlinear dynamics techniques and simulations that allow one to rank randomly growing trees for the data storage problem [EB,MH].
• Verify the instanton method on long LDPC codes and for different physical channels. Analyze scaling with the code size and the number of iterations. Develop new principles of coding design based on the recently drawn analogy between the theory of percolation and coding [MC,AH,ER,MS,ZT].
• Develop a classification of problems in verification & validation whose phase structure enables survey propagation to perform effectively. Analyze phase transitions in the Graph Bipartition problem [FA,JB,LG,GI,AP].
• Develop a cluster code for clique identification on parallel ASC machines [BA,ZN]. Describe the community detection problem as an inference problem [MH].

**FY07:**

• Develop a set of algorithms based on integral representations for LDPC codes [MC,AH,MS].
• Develop a formalism to connect phase transitions and computational complexity [FA,JB,LG,GI,AP].
• Test the clique algorithm on unsolved graphs such as the Keller-6 graph (3361 nodes, 4.6M links) [BA,ZN]. Apply BP to the inference formulation of the community detection problem and test the results against competing algorithms [MH].
• Formulate the SAT-membrane problem and verify the theory against numerics [EB].

**FY08:**

• Utilize information gained in FY06-07 to design efficient error-correcting schemes. Seek a practical application of the instanton toolbox [MC,AH,MS].
• Extend survey propagation analysis to approximation algorithms. Apply to v& v problems identified for this purpose in FY06 [FA,JB,LG,GI,AP].
• Analyze SDA on typical systems of practical relevance, e.g. the *protein folding* problem. Apply the newly developed community detection algorithm to real-world problems [BA,MH,ZN].
• Theoretical analysis (phase transitions, scaling) of the rate equations for the random membrane-SAT [EB]. Characterization of the probability for being unsatisfied in the satisfied phase [MC,MS].
• Develop a prototype for model checking of multiagent simulations [FA,GI,AP].

# 6 Budget Justification

The total budget is roughly $\$1,229K/\$1,287K/\$1,352K$ for FY06/07/08. The M&S is estimated to be $\$70K/\$73K/\$77K$. This includes $\$45K$ per year for visitors, external collaborates, computers and travel. The funding includes 3 new postdocs (funded 0.3-0.5FTE each), $\$117K/\$123K/\$129K$. The funding is allocated among the organizations: T-Div $\$793K/\$829K/\$871K$, CCS-Div $\$382K$/ $\$401K$/$\$421K$, and X-Div $\$54K/\$57K/\$60K$.

# References


[1] Web-references to new technology releases relevant to this proposal are posted at **http://cnls.lanl.gov/~chertkov/alg.htm**.

[2] Gallager, R.G. *Low density parity check codes* (MIT Press, Cambridge, 1963).

[3] Berrou, C., et. al Proc. IEEE Conf. , 1993, Geneva; vol. 2, p.1064–70.

[4] Sourlas, N. Nature **339**, 693 (1989).

[5] Montanari, A. & Sourlas, N. *Eur. Phys. J.* B **18**, 107 (2000); ibid **23**, 121 (2001).

[6] Kirkpatrick, S. & Sherrington, D., Phys. Rev. Lett. **35**, 1792 (1975).

[7] Fisher, D.S. & Huse, D.A., Phys. Rev. Lett. **56**, 1601 (1986).

[8] Mezard, M., Parisi, G., Virasoro, M.A., *Spin glass theory and beyond* (World Scientific, 1987).

[9] Percus, A., Istrate. G, and Moore, C, eds. *Computational Complexity & Statistical Physics* (Oxford Press, 2005).

[10] Mezard, M., Parisi, G. & Zecchina, R., Science **297**, 812 (2002).

[11] Kadanoff, L., *Statistical Physics: Statistics, Dynamics and Renormalization*, (World Scientific, 2000).

[12] Bethe, H.A. Proc.Roy.Soc.London A, **150**, 552 (1935).

[13] Chertkov, M. et al. PNAS **98**, 14208 (2001); JOSA B 19, 2538 (2002); Phys. Rev. E.**67**, 036615 (2003); Chernyak, V., Chertkov, M., Gabitov, I., Kolokolov, I. & Lebedev, V. JLT **22**, 1155 (2004).

[14] Jain, V.K. et al., Fiber And Integrated Optics **20**, 95 (2001).

[15] Chernyak, V., Chertkov, M., Stepanov, M.G. & Vasic, B., Phys. Rev. Lett. **93**, 198702 (2004); Stepanov, M.G., Chernyak, V., Chertkov, M. & Vasic, B. submitted to Science.

[16] Ben-Naim, E., Krapivsky, P.L. & Majumdar, S. Phys. Rev. E **64**, R035101 (2001).

[17] Hastings, M.B. & Halsey, T.C. Europhys. Lett. **55**, 679 (2001).

[18] Gudkov, V., Nussinov, S. & Nussinov, Z. cond-mat/0209419.

[19] Boettcher, S & Percus, A, Artificial Intelligence **119**, 275 (2000).

[20] Knuth, D.E *The art of computer programming, vol. 3*, (Addison-Wesley, 1998).

[21] Wu, Y., et. al IEEE Trans. on Magnetics **36**, 2176 (2003).

[22] Koetter, R. & Vontobel, P.O. Proc. 3rd International Symposium on Turbo Codes & Related Topics, Symp. on Turbo codes, Brest, p. 75–82, 2003.

[23] Clarke, E. et al. *Model Checking* (MIT Press, 2000).

[24] C. Moore, G. Istrate, D. Demopoulos, M. Vardi, http://arxiv.org/abs/math.PR/0505032 .

[25] Vazirani, V. *Approximation Algorithms* (Springer, 2001).

[26] J. R. Banavar, D. Sherrington, and N. Sourlas, J. Phys. A **20**, L1 (1987).

[27] Istrate, G., Boettcher, S. & Percus, A., http://arxiv.org/abs/cs.CC/0503082.

[28] Bollobas, B. et al., Rand. Struc. Alg. **18**, 201 (2001).

[29] Ben-Naim, E. & Krapivsky, P.L., Phys. Rev. E **71**, 026129 (2005).

[30] Carraghan, R., & Paradalos, P.M. Oper.Res.Lett. **9**, 375 (1990).

[31] Girvan,M. & Newman, M. E. J. PNAS **99**, 7821 (2002).

# A Overview of key Participants

A key element in our strategy is assembling a very broad but highly expert team. The expertise of the individuals contributing to this project has been acquired through their previous experience, but also results from ongoing research funded by programmatic efforts and current LDRD projects. One such project is the FY04-FY06 DR on epidemiological and infrastructure networks. The two projects address two very distinct problems. Some of the graph theoretical capabilities developed in this former and ongoing projects will be utilized and applied for advancing research on coding theory and combinatorial optimization proposed in the present DR. We stress that this team has shown it can work well together. This proposal allows LANL to solidify its critical mass in statistical physics for national security applications.

In the past 3 years, the top 6 participants in this proposal have authored 3 Nature articles, 3 PNAS articles, 1 Science article, and 1 US patent. In that period they authored a total of over 120 publications in peer-reviewed journals including over 50 in Physical Review, and over 25 in Physical Review Letters. They have also edited 3 books. Visit **http://cnls.lanl.gov/~chertkov** for detailed CVs and the full list of relevant references.

## A.1 Key Participants' Role:

Chertkov, Ben-Naim, Hastings, Stepanov, and Toroczkai are responsible for developing unifying theoretical physics inspired approaches to analysis and improvement of algorithms. This sub-team will be working across the problems to cross-fertilize the theoretical physics ideas central for the project success. The external collaborators Chernyak, Krapivsky, and Mezard, all internationally known physicists, will help with the theoretical methods development. The Lead Co-PI on this is Chertkov.

Istrate, Alexander, Gurvits, Percus are responsible for developing computational complexity theory approaches. Moore is an external collaborator helping on the mathematical aspects of the project. The Lead Co-PI on this is Istrate.

Hansson, Ben-Naim, Chertkov, Ravasz, Stepanov, and Toroczkai are responsible for developing novel coding schemes. Koetter and Vasic are external collaborators bringing outstanding expertise in coding-theory and electrical-engineering to the project. The Lead Co-PI on this is Hansson.

Percus, Barré, Ben-Naim, Chertkov, Hastings, Istrate, and Stepanov will be developing heuristic approaches to evaluation, analysis and improvement of algorithms for satisfiability. The Lead Co-PI on this is Percus.

Hastings, Albright, and Nussinov will be developing novel physics approaches to difficult problems in community detection. The Lead Co-PI is Hastings.

Ben-Naim, Albraight, Alexander, Chertkov, Percus, Stepanov, Toroczkai, Ravasz have the role of performing large-scale distributed simulations in data clustering, optimization and error-analysis. The Lead Co-Pi on this is Ben-Naim.

## A.2 Brief Bio of Key Participants:

**Michael Chertkov** (PI, T-13, 0.75FTE) received his Ph.D. in theoretical physics in 1996 from Weizmann Institute. For his seminal Ph.D. thesis on anomalous scaling in scalar turbulence, he received the Prize of the Charles Clore Israel Foundation in 1995 given to the best Ph.D. student this year in Physics in Israel. He was an R. H. Dicke Fellow at Princeton University, Physics Department and moved to LANL in 1999 as J.R. Oppenheimer Fellow in T-13/CNLS and joined

T-13 in 2002 as a Staff Member. His area of expertise includes theory of turbulence, statistical and nonlinear optics, information theory, error-correction theory, and soft condensed matter. He has one patent and over 50 publications in refereed journals including PNAS, and 11 publications in Physical Review Letters. He organized 5 CNLS conferences, most recent in January 2005 on *Statistical Physics approach to Coding Theory*. His achievements most relevant to this project are (i) development of novel physics inspired approach which ultimately solves the problem of the error-floor analysis of the coding theory (co-authored with M. Stepanov, V. Chernyak and B. Vasic) and (2) comprehensive statistical analysis of effects of noise and disorder on pulse propagation in fiber optics resulted in 13 publications in optics, physics and engineering journals.

**Brian Albright** (X-1, 0.25FTE) received his Ph.D. in physics from UCLA in 1998 for work on statistical properties of topological singularities in random vector fields and implications for the formation of finite-time current singularities. This work was funded in part by a Fletcher Jones Foundation fellowship. Albright was a postdoc at UCLA working on thermal transport in stochastic plasma media, and then joined X-1 as a postdoc and then TSM. Albright contributes extensive experience with the development and implementation of numerical algorithms in high performance computing settings, and he has actively participated in the development of the SDA algorithm. He is a co-inventor of Quiet Monte Carlo direct simulation, and he maintains the LANL VPIC particle-in-cell (PIC) simulation code, the fastest (by an order of magnitude) fully-relativistic 3D PIC simulation code extant. Albright has 48 publications in the open and classified literature.

**Frank Alexander** (CCS, 0.1FTE) received Ph.D. in Physics from Rutgers U in 1991. He was a post-doctoral fellow at CNLS/LANL in 1991-1993 and at LNL in 1993-1995. At Livermore he developed algorithms for computational kinetic theory and hybrid numerical methods. From 1995 to 1998 Frank was a research assistant professor at the Center for Computational Science at Boston University. While in Boston, he developed methods to improve computation in nonequilibrium physics. In 1998 Frank moved back to Los Alamos as a staff member in CIC-19 (now CCS-3). Since his return, he has been working on a variety of problems in the above areas as well as in applying methods of statistical physics to modeling complex systems and time series analysis. Frank became team leader in 2000 and he is Deputy Group Leader since 2002.

**Julien Barré** (CNLS/T-11) received his PhD in theoretical physics in 2003 from the École Normale Supérieure de Lyon (France) and the University of Florence (Italy). His research focused on statistical mechanics and non linear dynamics of long range interacting systems. As a director's funded postdoc in LANL since Fall 2003, he worked among other themes on disordered systems and percolation transitions. He has 13 refereed publications, including 3 in Physical Review Letters, and his knowledge in methods of disordered systems are relevant for the proposed research.

**Eli Ben-Naim** (T-13, 0.75FTE) received his Ph.D. in Physics from Boston University in 1994, was postdoctoral research associate at the University of Chicago (1994-1996), Director's postdoctoral fellow at Los Alamos (1996-1998), and a technical staff member at the Complex Systems group since. He is an expert in statistical physics, nolinear physics, and random processes with fundamental contributions to the kinetic theory of granular materials and traffic flows. His recent works on kinetic theory of random structures including random graphs and random trees are directly relevant for the proposed research. He authored over 75 publications in peer-reviewed journals cited over 1200 times. He serves on the editorial board of Physical Review E (the leading statistical and nonlinear physics journal) and the advisory board of Journal Physics A.

**Leonid Gurvits** (CCS-3, 0.1FTE) is an expert in theoretical computer science, control theory and quantum information. He published in Advances in Mathematics, Journal of Computer and System Sciences, SIAM Journal on Matrix Analysis, Foundations of Computational Mathematics and Physics Review, and was multiple times an invited speaker at the most selective theoretical computer science conferences including STOC (theory of algorithms).

**Anders Hansson** (CCS-DO,0.5FTE) received his Ph.D. in communication theory from Chalmers University of Technology, Göteborg, Sweden, in 2003. He spent the 2000-2001 academic year as Sweden-America Foundation Fellow at the Communication Sciences Institute of USC. During 2003-2005, he was a Postdoctoral Research Associate at LANL, and since January 2005, he is a Technical Staff Member in CCS-DO. Dr. Hansson's area of expertise includes spatio-temporal fading, receiver front-end processing, continuous phase modulation, and adaptive soft-input soft-output algorithms. Parts of his research have been implemented by TrellisWare Technologies, San Diego, CA, for deployment in U.S. government communication systems. His recent interests include simulation and modeling of large- scale socio-technical systems, acceleration of high-performance computing algorithms through reconfigurable supercomputing, and sequential dynamical systems. He published 25 papers including four papers in IEEE Trans. on Communications.

**Matthew Hastings** (T-13, 0.75FTE) received his Ph.D. in theoretical physics in 1997 from the Massachusetts Institute of Technology. His seminal Ph.D. thesis developing the conformal mapping algorithm for diffusion-limited aggregation led to his receiving the Lockett Award given to the best Ph.D. work in theoretical physics at MIT. He was an R. H. Dicke Fellow at Princeton University and then an R. P. Feynman Fellow in T-13/CNLS. He joined T-13 as a staff member in 2004. His area of expertise includes statistical physics, networks, non-equilibrium dynamics, strongly correlated electrons, and soft condensed matter. He has published 13 papers in Physical Review Letters and his work has been featured in Nature News and Views, Physical Review Focus, the Bell Labs Condensed Matter Journal Club, and Technology Research News. He received the LANL postdoctoral distinguished award for his work on Ratchet Cellular Automata.

**Gabriel Istrate** (CCS-DO, 0.5FTE) received his Ph.D. in Computer Science from the University of Rochester in 1999. His Ph.D. thesis deals with exactly-solved models of phase transition in combinatorial problems. He moved to LANL in 1999 as a Director Funded Postdoc, and joined CCS-DO (then TSA-SA) in 2001. His area of expertise includes computational complexity, probabilistic analysis of algorithms, combinatorial optimization and simulation of sociotechnical systems. He has authored 27 papers, 10 of which dealing with phase transitions in combinatorial problems. He served on the Program Committee for ICALP (the most prestigious European conference in Theoretical Computer Science) in 2005 and is the editor (together with Allon Percus and Cris Moore) of a volume, "Computational Complexity and Statistical Physics", scheduled to appear at Oxford University Press in June 2005.

**Zohar Nussinov** (T-11, 0.5FTE in FY06) PhD UCLA 2000. He is an author of more than 40 works including publications in Nature and Physical Review Letters. He has considerable experience with the theory and modeling of glasses, related thermodynamics, and complex systems. In his thesis, he developed an approach to understanding the glass transition in supercooled liquids based on notions of geometrically constrained dynamics in low dimensions. This led to a remarkably successful universal fit for the relaxation times in all known glass formers (a relatively cited work- with more than 150 citations). Extensions of this idea led to the Simplex Distortion Algorithm (SDA) which will be studied in this proposal.

**Allon Percus** (CCS-3, 0.5FTE) received his Ph.D. from Orsay, in 1997. In addition to his LANL duties, he is currently Associate Director of the Institute for Pure and Applied Mathematics at UCLA. His main research interests are in discrete optimization and statistical physics. Much of his work is on the interface between these two fields. In his research on the stochastic traveling salesman problem, he produced the most precise numerical estimate to date for large $n$ asymptotic tour lengths. He was a co-founder of the Extremal Optimization method, which has been used successfully on several hard combinatorial optimization problems. He is PI on the LDRD/ER *Improving Local Search*, and has also led past LDRD efforts in *Extremal Optimization* and *Combinatorial Optimization in Biology*. He has organized numerous conferences and workshops on combinatorics, phase transitions and algorithmic complexity, including a symposium on *Phase Transitions in Computer Science* at the 2004 AAAS Annual Meeting that highlighted recent breakthroughs in statistical physics, computational threshold phenomena and coding theory.

**Mikhail Stepanov** (CNLS/T-13, 0.75 FTE) was “Soros” graduate student for three years in a row and received Ph.D. in physics in 1999 from Institute of Automation and Electrometry, Russia. He was a postdoc in Weizmann Institute 1999-2001, member in the Institute for Advanced Studies, Princeton 2002-2003, and joined CNLS/T-13 as a postdoc in 2004. His area of expertise include nonlinear spectroscopy, fiber optics, turbulence, pattern growth, data clustering and error correcting codes. He has 17 publications in refereed journals, including Nature, and 3 publication in Phys. Rev. Lett. His expertise most relevant to this project is in numerical analysis of clustering algorithms and instantons in turbulence and error-correction.

**Erzsébet Ravasz** (CNLS/T-13) is a director-funded postdoctoral fellow. She received her PhD in 2000 from Notre Dame, working with A.-L. Barabasi. Her expertise is in community detection, networks, and synchronization. She has published 12 papers, including papers in Nature and Science, with over 450 citations.

**Zoltán Toroczkai** (CNLS, 0.2FTE in FY06) Zoltan Toroczkai is the Deputy Leader of the CNLS, and has been a staff member at the Complex Systems Group since June 2002. He was a Director’s Postdoctoral Fellow in the period 2000-2002. He received his PhD in Theoretical Physics from Virginia Tech in 1997 with specialization in nonequilibrium statistical physics. He is the author of over 50 publications, including 2 Nature articles (on networks), 1 Science article (on parallel computing) 2 PNAS article (on chemical reactions and mixing) and 1 article in Physics Today (surface growth). His areas of expertise include complex networks, agent-based systems, statistical mechanics, fluid flow, dynamical systems and chaos theory.

## A.3 External Collaborators:

Our advisory board consists of the following six external collaborators. The board will convene yearly to review progress and plan future work. These collaborators are all internationally well-known scientists and they cover a broad range of expertise.

**Vladimir Chernyak** is Professor at Wyane State University. He is the world leading expert in statistical field theory, quantum optics, physical chemistry and fiber optics, who published more then 180 papers cited more then 2,600 times total. He has four years of industrial experience in optics communication as Senior Research Scientist at Corning Inc.

**Ralf Koetter** is Professor at the Coordinated Science Laboratory at the UIUC. In 2000, he started a term as associate editor for coding theory of the IEEE Transactions on Information Theory. He received an IBM Invention Achievement Award in 1997 and an IBM Partnership Award in 2001. He is a member of the Board of Governers of the IEEE Information Theory Society. He was recently awarded the 2004 Best Paper Award by the IEEE Information Theory Society. For further information see: http://ww.comm.csl.uiuc.edu/˜koetter .

**Paul Krapivsky**, Professor at BU, is an internationally recognized figure in nonequilibrium statistical physics with over 150 publications in the past 10 years. His recent work on the theory of random growing structures is seminal. His expertise relevant to this proposal includes rate equation analysis of data structure and phase transitions in computational complexity.

**Marc Mezard** is a 'directeur de Rrecherche' in CNRS (Orsay). He has received the golden medal of CNRS in 1990 and the Ampere prize of the french academy of science in 1996. He has been PI of a European network of 13 European laboratories on "Statistical physics of collective behaviour in disordered systems and information processing" in 1993-1997 and is presently the PI of a European network of 10 laboratories on "Statistical physics of information processing and combinatorial optimization". He is the inventor of the *survey propagation* algorithm.

**Cristopher Moore** is Professor at UNM. He has over 80 publications. Recent work relevant to this proposal includes the structure of the internet, the phase transition in satisfiability, and the computational complexity of physical simulations.

**Bane Vasic** is Professor at the Electrical and Computer Engineering Department at UA, Tucson. In the past he worked at Kodak Research and Bell Laboratories, where he was involved in research on optical storage as well as development of codes and detectors implemented in chips. He holds nine US patents. He is a Member of the Editorial Board of the IEEE Trans. on Magnetics.

# B Time allocation

Time commitments are listed below for current projects and proposals submitted for this year.

**Michael Chertkov (T-13)** is currently funded by Instabilities and T-Mix (0.5FTE) and LDRD DR proposal "Lagrangian Measurements ..." (0.5FTE). Future possible involvements: He is on LDRD ER proposal "Novel physics ..." (0.3FTE) and on LDRD ER proposal "Suppression of ..." (0.2FTE). In case this DR is funded he will adjust the time on the rest of the projects such that he will be funded on the 0.75FTE level on this project.

**Brian Albright (X-1)** will be funded in FY06 at the 0.5 FTE level on one LDRD (20040064DR; PI: Fernández) and at the 0.25 FTE level on NASA LWS TR&T funding. Albright is PI on a pending LDRD ER (20060453ER; 0.33 FTE) and Co-PI on another LDRD DR (20060060DR; PI: Reeves). Any time not defined here will be covered by programmatic sources; adjustments will be made if more support is awarded from these sources than can be spent.

**Frank Alexander (CCS-3)** funding will be split 0.1 FTE/0.9FTE between this LDRD-DR and Overhead (Deputy GL duties).

**Julien Barré (CNLS/T-11)** is a director funded postdoctoral fellow.

**Eli Ben-Naim (T-13)** is currently funded by the LDRD-DR project "statistical physics of infrastructure networks". He submitted one ER proposal "Energy distributions in granular flows" (0.5FTE). If this proposal is funded, he will adjust the time on the rest of the projects such that he will be funded on the 0.5FTE/0.75 FTE level on this project in FY06/FY07-08.

**Leonid Gurvits (CCS-3)** is currently funded by the LDRD-ER project “Classical complexity ...” (0.5 FTE) and LDRD-DR project “Physics of Information” (0.5FTE). He is participant on an ER and a DR submitted this spring. He will be funded on the 0.1FTE level on this project and will adjust the time on the rest of the projects accordingly.

**Anders Hansson (CCS-DO)** is currently funded by three projects: ”Scalable and Reconfigurable Computing” (0.5FTE), ”Generic Cities” (0.25FTE), and ”AdHop-Net” (0.25FTE). His level of commitment in the two last projects is flexible at this point. Future possible involvement includes two LDRD-ER proposals: ”Specification of Multiagent Systems” (0.4FTE), and ”A Calculus for Gene-Regulatory Networks” (0.2FTE). In case this DR is approved he will be funded on the 0.5FTE level on this project and will adjust the time on the rest of the projects accordingly.

**Matthew Hastings (T-13)**. If the project is funded he will work 50% on this project and 50% on LDRD-DR “Statistical Physics of Infrastructure Networks” in FY06, and will work 75% on this project in future years.

**Gabriel Istrate (CCS-DO)** is currently funded by the AdHopNet project (0.5 FTE), by the LDRD-DR “Statistical Physics of Infrastructure Networks” (0.25 FTE) and the LDRD-ER “Local Search” proposal (0.25 FTE). He is involved in the submission of two ER proposals. In case this DR is funded he will adjust the time on the rest of the projects such that he will be funded on the 0.5FTE level on this project.

**Zohar Nussinov (T-11)** is a postdoctoral researcher currently fully funded by LDRD 20030036DR (X1WX, PI: Joe Thompson). In case this DR is approved he will be funded on the 0.5FTE level on this project in FY06.

**Allon Percus (CCS-3)** is currently funded by the LDRD-ER “Local Search” (0.35 FTE), the LDRD-DR “Statistical Physics and Infrastructure Networks” (0.15 FTE) and the UCLA Institute for Pure and Applied Mathematics (0.5 FTE). He is involved in the LDRD-ER proposal “Optimization in Percolation” submitted for the FY06 funding round, with a tentative commitment of 0.3 FTE. He intends to devote 0.5FTE to the project proposed here, and will be able adjust the exact level of commitments on other projects as needed.

**Erzsébet Ravasz (CNLS/T-13)** is director funded postdoctoral fellow, thus fully funded.

**Mikhail Stepanov (CNLS/T-13)** is currently postdoctoral researcher funded by LDRD ER “Secure communications ...” (0.5 FTE) and by CNLS/LDRD grant (0.5 FTE). In case this DR is funded he will adjust the time on the rest of the projects such that he will be funded on the 0.75 FTE level on this project.

**Zoltán Toroczkai (CNLS)** is currently funded on 0.75FTE level by CNLS and 0.25FTE by the LDRD DR ”Statistical Physics of Infrastructure Networks”. He will adjust his time on CNLS to accommodate the 0.2FTE on this project.